Wstęp do probabilistyki i statystyki

Wykład 1.

Wstęp

dr hab.inż. Katarzyna Zakrzewska, prof.AGH
Katedra Elektroniki, AGH
e-mail: zak@agh.edu.pl
http://home.agh.edu.pl/~zak

Literatura:

- D.C. Montgomery, G.C. Runger, *Applied Statistics and Probability for Engineers*, J. Wiley and Sons, Inc.
- A. Plucińska, E. Pluciński, Probabilistyka, rachunek prawdopodobieństwa, statystyka matematyczna, procesy stochastyczne, WNT, 2000
- J. Jakubowski, R. Sztencel, Wstęp do teorii prawdopodobieństwa, SCRIPT, 2000
- R. Leitner, W. Żakowski, Matematyka-kurs przygotowawczy na wyższe uczelnie techniczne

Plan:

- Rys historyczny
- Rodzaje danych
- Prezentacja danych
- Zastosowania statystyki
- Parametry opisowe

Czy zajmuje się probabilistyka i statystyka?

Teoria prawdopodobieństwa (także **rachunek prawdopodobieństwa** lub **probabilistyka**) – dział matematyki zajmujący się zdarzeniami losowymi. Zdarzenie losowe to wynik *doświadczenia losowego*.

Doświadczenie losowe może być powtarzane dowolnie wiele razy w warunkach identycznych lub bardzo zbliżonych a jego wynik nie daje się przewidzieć jednoznacznie.

Częstość zdarzenia:

$$\frac{l}{n}$$

l – oznacza ile razy zaszło dane zdarzenie gdy doświadczenie powtarzano n razy

Prawidłowość statystyczna – przy coraz większej liczbie doświadczeń losowych częstość zdarzenia dąży do pewnej stałej liczby

Czy zajmuje się probabilistyka i statystyka?

- Rachunek prawdopodobieństwa zajmuje się badaniem abstrakcyjnych pojęć matematycznych stworzonych do opisu zjawisk, które nie są deterministyczne:

1. zmiennych losowych w przypadku pojedynczych zdarzeń oraz
2. procesów stochastycznych w przypadku zdarzeń powtarzających się (w czasie).

- Jako matematyczny fundament statystyki, teoria prawdopodobieństwa odgrywa istotną rolę w sytuacjach, w których konieczna jest analiza dużych zbiorów danych.

Jednym z największych osiągnięć fizyki dwudziestego wieku było odkrycie probabilistycznej natury zjawisk fizycznych w skali mikroskopowej, co zaowocowało powstaniem mechaniki kwantowej.

Statystyka zajmuje się metodami zbierania informacji (liczbowych) oraz ich analizą i interpretacją.

Rys historyczny

- Matematyczna teoria prawdopodobieństwa sięga swoimi korzeniami do analizy gier losowych podjętej w siedemnastym wieku przez Pierre de Fermata oraz Blaise Pascala.
- Z tego powodu, początkowo teoria prawdopodobieństwa zajmowała się niemal wyłącznie zjawiskami dyskretnymi i używała metod kombinatorycznych. Zmienne ciągłe zostały wprowadzone do teorii prawdopodobieństwa znacznie później.
- Za początek stworzenia współczesnej teorii prawdopodobieństwa powszechnie uważa się jej aksjomatyzację, której w 1933 dokonał Andriej Kołmogorow.

Rys historyczny

Blaise Pascal (1601-1662)





XVII w. , Paryż, Francja

Unieśmiertelnił kawalera de Méré oraz jego paradoks hazardowy prawdopodobieństwa wyrzucenia szóstek na jednej i dwóch kościach.

„Trójkąt Pascala" wykorzystywany przy potędze sumy

Rys historyczny

Początek XVII w., Touluse, Francja

Pierre de Fermat (1601-1665)



Badał właściwości liczb pierwszych, teorię liczb, równolegle opracował metodę współrzędnych w geometrii. Razem z Pascalem stworzył podstawy pod współczesny rachunek prawdopodobieństwa.

Rys historyczny

XVIII-XIX w., Paryż, Francja

Siméon-Denis Poisson

(1781-1840)



Przyjaciel Lagrange'a, uczeń Laplace'a na sławnej École Polytechnique.

Poza zagadnieniami mechaniczno -fizycznymi zajmował się teorią prawdopodobieństwa.

Proces stochastyczny (podobnie jak pr. Markowa), rozkład Poissona - dystrybuanta!

Rys historyczny

Carl Frederich Gauss (1777-1855)



XVIII-XIX w., Getynga, Niemcy

Profesor Universytetu w Getyndze

Genialny matematyk, który już w dzieciństwie wyprzedzał umiejętnościami rówieśników (w szkole podstawowej jako jedyny rozwiązał zadanie nauczyciela - zsumowanie liczb 1 do 40 – zauważając że jest to (40+1)*20)

Rozkład normalny, zwany krzywą Gaussa

Paradoks kawalera de Méré

De Méré, zapalony gracz w kości, dokonał obserwacji, że częściej wypada jedna szóstka przy 4 rzutach jedną kostką niż dwie szóstki przy 24 rzutach dwiema kostkami.

Wg. (błędnej) logiki gracza:

Na jednej kostce: $4 * 1/6 = 4/6$

Na dwóch kostkach: $24 * 1/6 * 1/6 = 24/36 = 4/6$

Zdarzenia wydają się mieć takie samo prawdopodobieństwo, dlaczego zatem hazardzista obserwował inny wynik?

Rozwiązanie paradoksu de Méré

Prawidłowo obliczone prawdopodobieństwo owych zdarzeń:

a) wyrzucenie co najmniej jednej szóstki przy 4 rzutach jedną kostką = 1 – prawdopod. nie wyrzucenia żadnej szóstki przy 4 rzutach jedną kostką =

$$1-\left(\frac{5}{6}\right)^4=\frac{671}{1296}\approx 0{,}5177$$

b) wyrzucenie co najmniej raz dwóch szóstek przy 24 rzutach dwiema kostkami = 1- prawdopod. niewyrzucenia dwóch szóstek przy 24 rzutach 2 kostkami =

$$1-\left(\frac{35}{36}\right)^{24}\approx 0{,}4914$$

STATYSTYKA

OPISOWA ANALIZA DANYCH

(DESCRIPTIVE STATISTICS)

- ***Organizacja danych***
- ***Podsumowanie danych***
- ***Prezentacja danych***

GRAFICZNA

NUMERYCZNA

DEDUKCYJNA – MODELOWANIE STOCHASTYCZNE

(STATISTICAL INFERENCE)

Podaje metody formułowania wniosków dotyczące obiektu badań (populacji generalnej) w oparciu o mniej liczny zbiór (próbę)

Typy danych

ILOŚCIOWE

(QUANTITATIVE, NUMERICAL)

Przykłady:

- ***Zbiór ludzi***
- ***Wiek***
- ***Wzrost***
- ***Wysokość zarobków***

Obliczenia pewnych parametrów, jak np. średnia arytmetryczna, mediana, ekstrema, mają sens

JAKOŚCIOWE

(QUALITATIVE, CATEGORIAL)

Przykłady:

Płeć

Stan cywilny

Można przypisać różnym cechom arbitralne wartości liczbowe.

Obliczenia parametrów nie mają sensu, można jedynie podawać np. udział procentowy

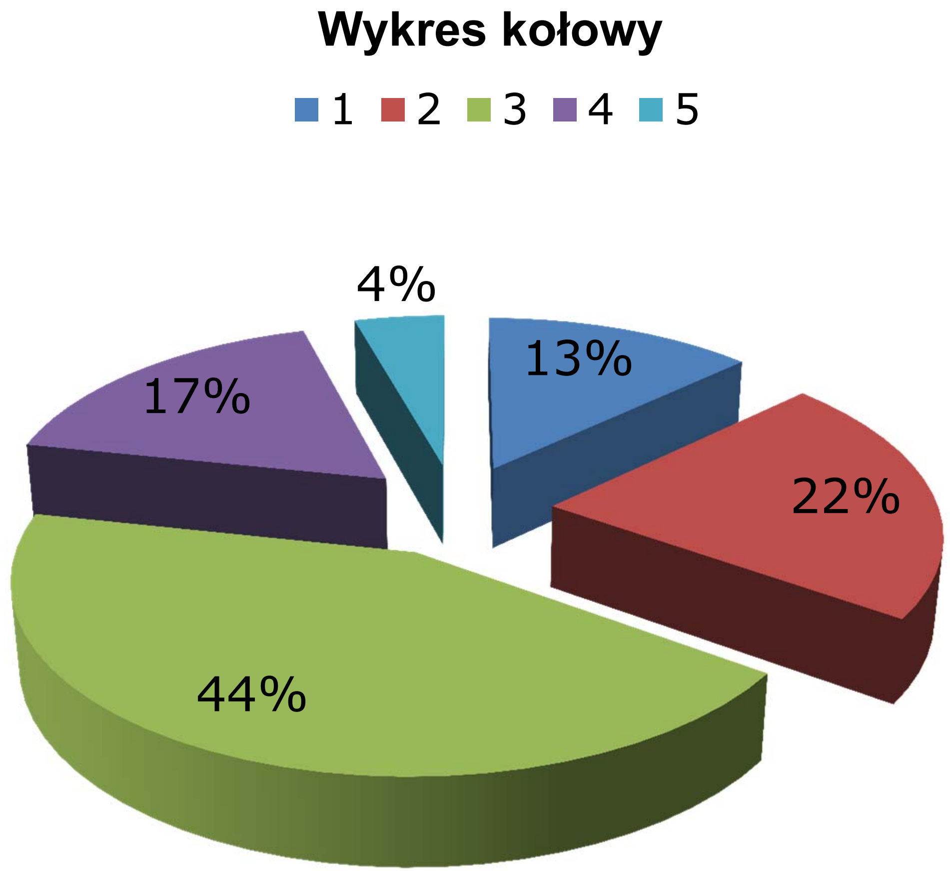
Graficzna prezentacja danych

x	Ilość wystąpień	Częstość względna
1	3	3/23 = 0,1304
2	5	5/23 = 0,2174
3	10	10/23 = 0,4348
4	4	4/23 = 0,1739
5	1	1/23 = 0,0435
Razem:	23	1,0000

Dane jakościowe można prezentować na wiele sposobów żeby zobrazować np. częstość występowania danej cechy

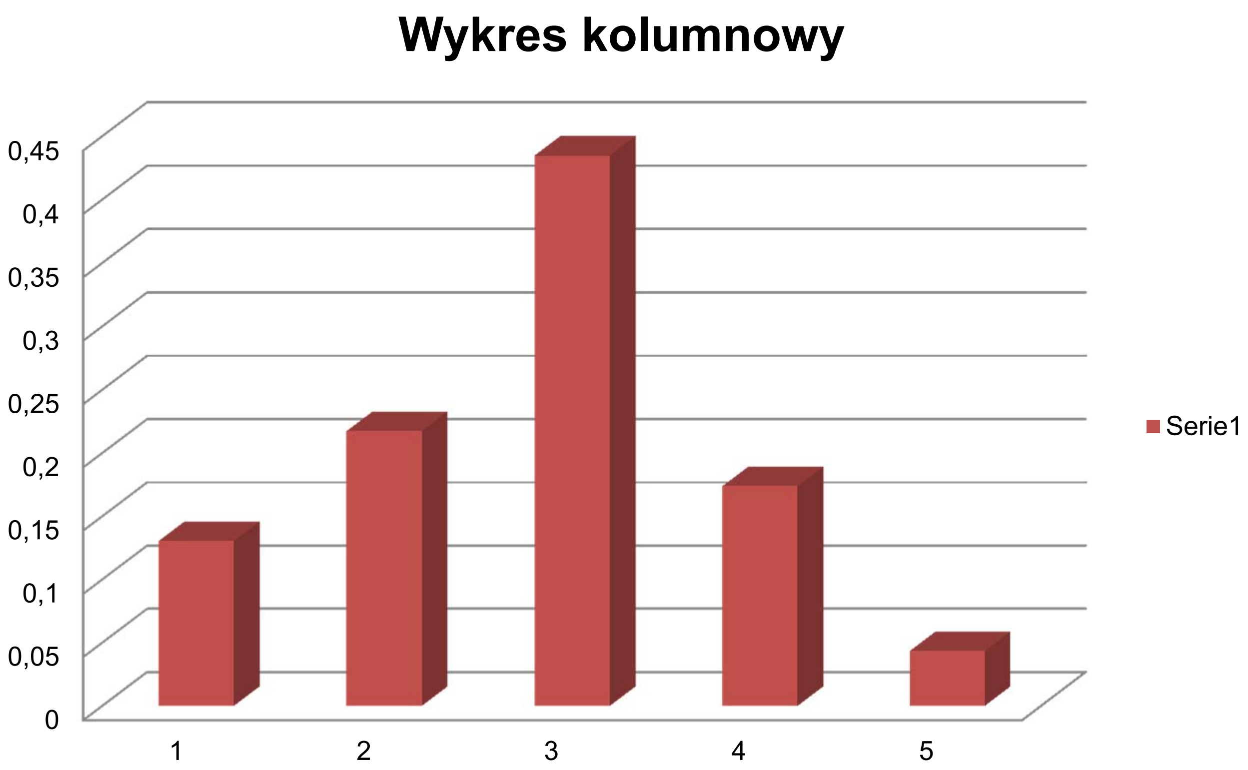
Graficzna prezentacja danych

1	0,13043478
2	0,2173913
3	0,43
4	0,17391
5	0,04347826



Graficzna prezentacja danych

1	0,13043478
2	0,2173913
3	0,43
4	0,17391
5	0,04347826



Dane ilościowe

Wyniki 34 pomiarów (np. wielkość ziaren w [nm], temperatura w kolejnych dniach o godz. 11:00 w [deg. C], czas rozmów telefonicznych w [min], itp.

3,6	13,2	12	12,8	13,5	15,2	4,8
12,3	9,1	16,6	15,3	11,7	6,2	9,4
6,2	6,2	15,3	8	8,2	6,2	6,3
12,1	8,4	14,5	16,6	19,3	15,3	19,2
6,5	10,4	11,2	7,2	6,2	2,3	

Tak podane wartości są mało czytelne!

Histogram

Sporządzenie wykresu (histogramu):

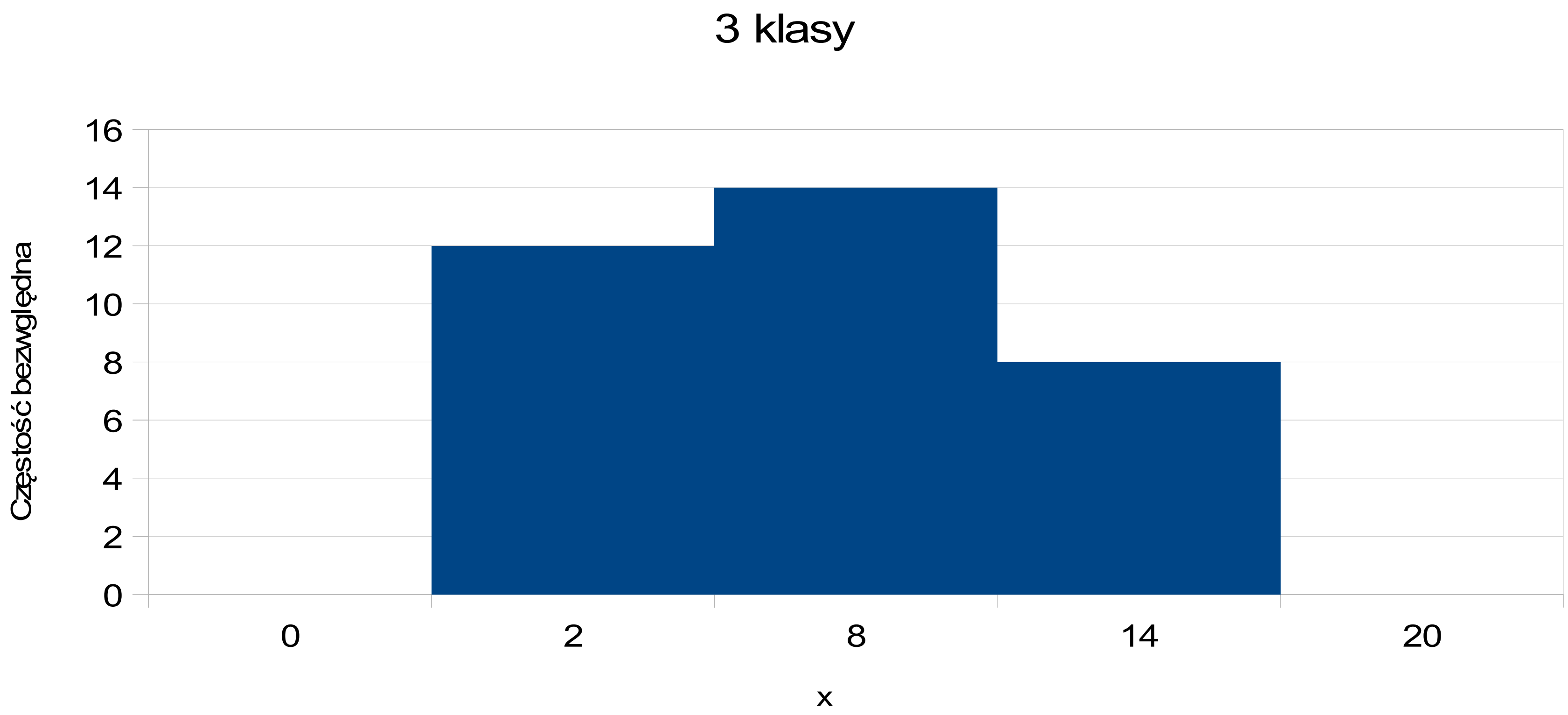
1. Uporządkować zbiór wg. rosnących (lub malejących) wartości – program Excel ma taką opcję.

2. Wyniki próby (o liczebności n) stanowią zbiór n-liczb (niekoniecznie różniących się od siebie). Celem ich ilustracji dzieli się je na klasy, tworząc tzn. szereg rozdzielczy.

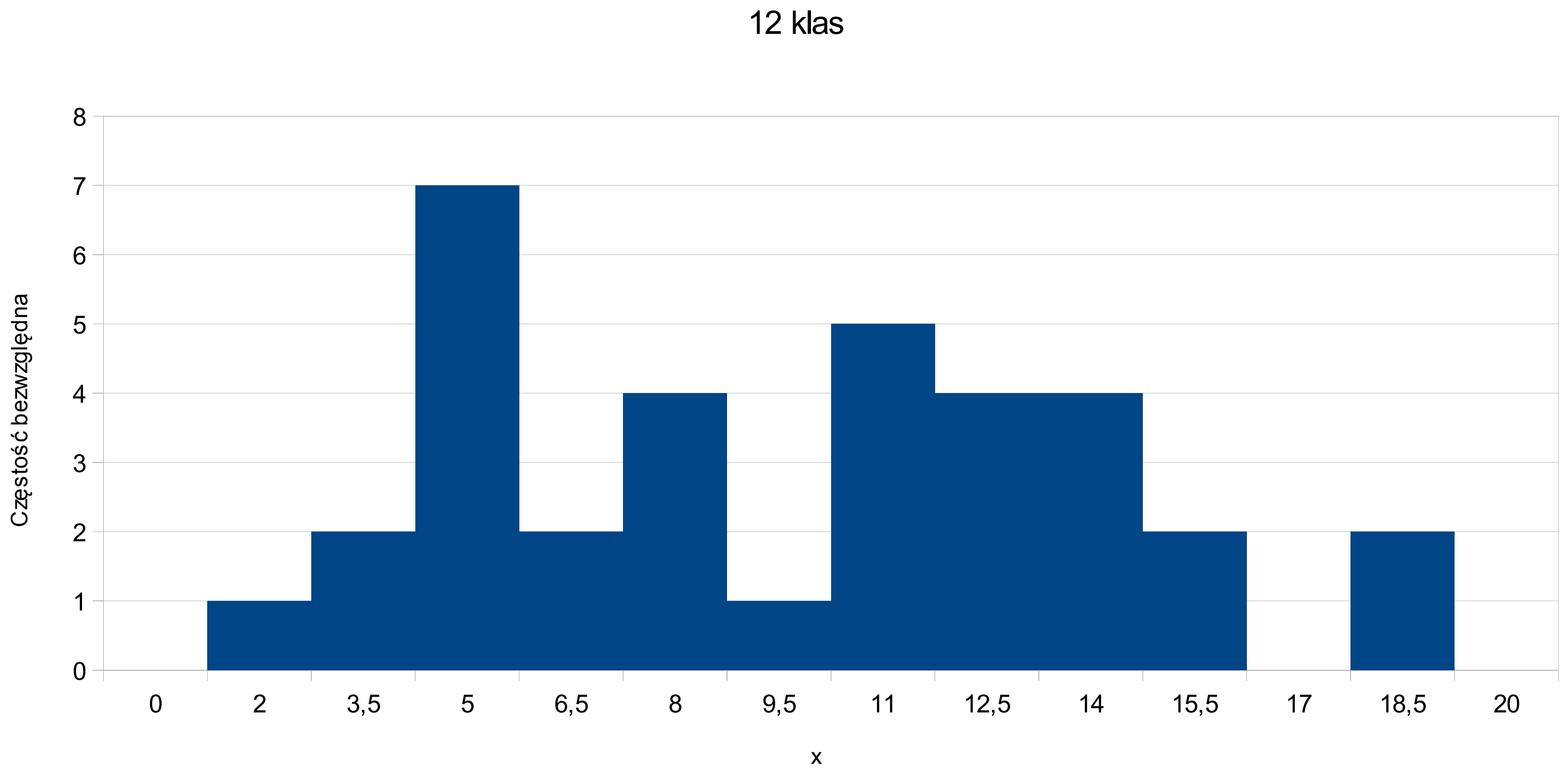
3. Szerokość poszczególnych klas nie musi być taka sama, choć zwykle stosuje się klasy o tej samej szerokości

4. Ilość klas nie może być zbyt mała ani też zbyt liczna. Najbardziej optymalną liczbę klas 'k' określa reguła Sturge'a.

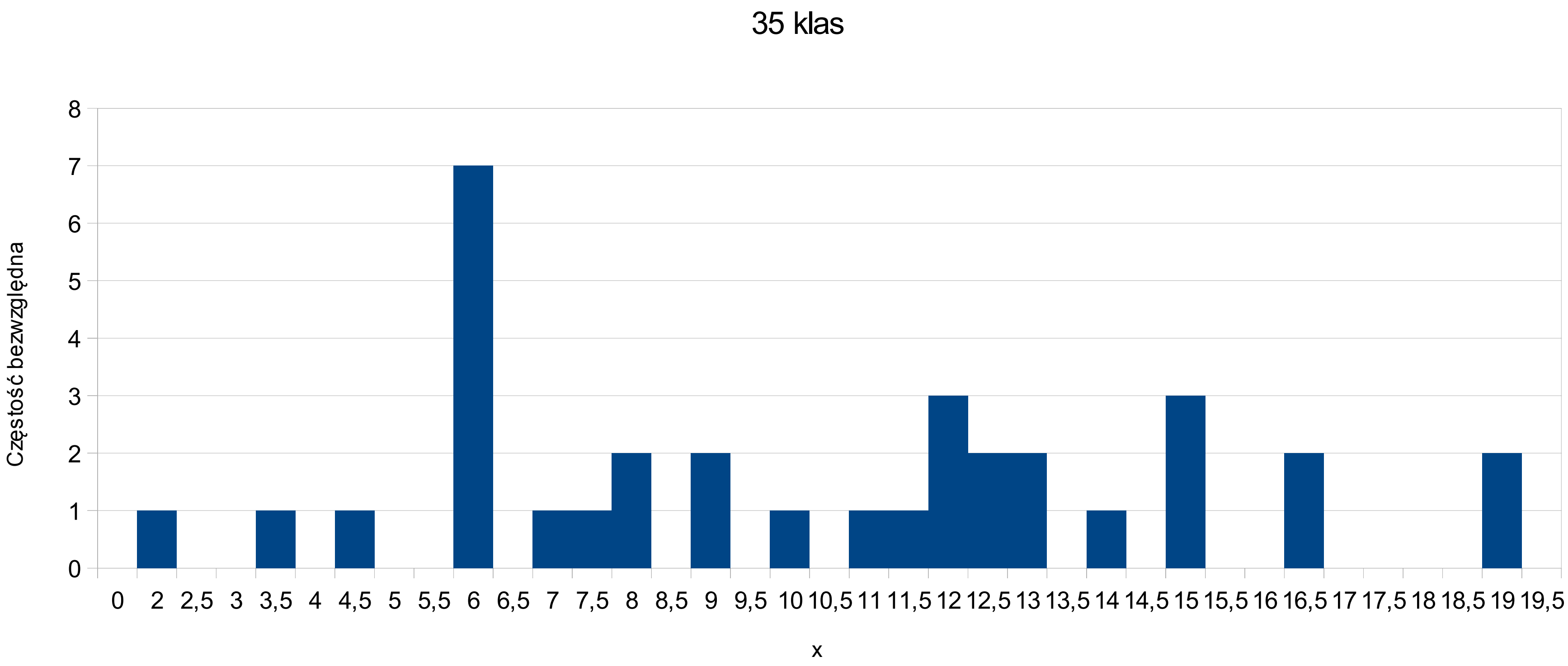
Histogram



Histogram


12 klas
Częstość bezwzględna
0
1
2
3
4
5
6
7
8
0
2
3,5
5
6,5
8
9,5
11
12,5
14
15,5
17
18,5
20
x

Histogram


35 klas
Częstość bezwzględna
8
7
6
5
4
3
2
1
0
0 2 2,5 3 3,5 4 4,5 5 5,5 6 6,5 7 7,5 8 8,5 9 9,5 10 10,5 11 11,5 12 12,5 13 13,5 14 14,5 15 15,5 16 16,5 17 17,5 18 18,5 19 19,5
x

Reguła Sturge'a

$$k = 1 + 3{,}3\log_{10} n$$

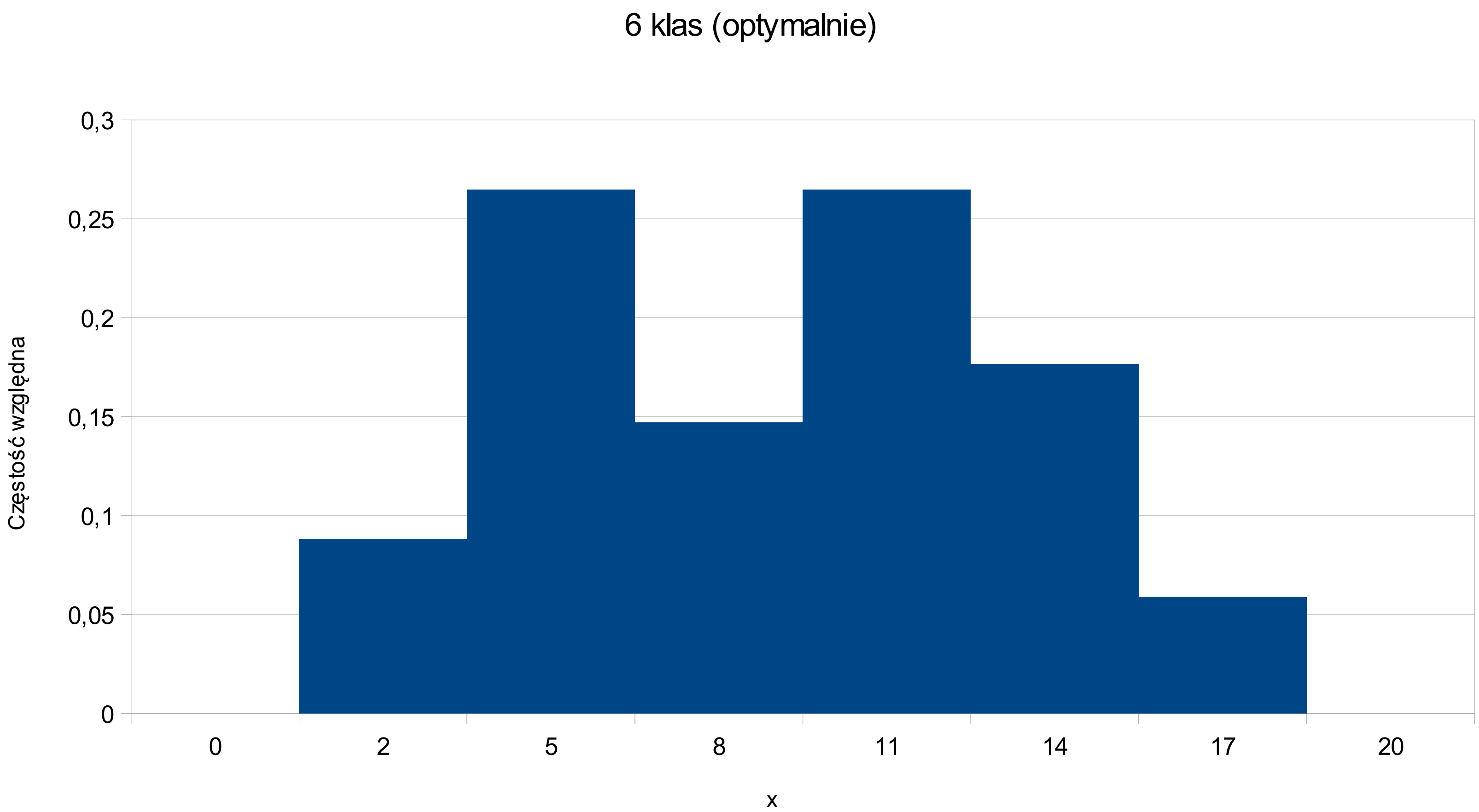
Dla naszego przykładu:

$$n = 34$$

$$k = 5.59 \approx 6$$

Liczebność próbki, n	Liczba klas, k
< 50	5 – 7
50 – 200	7 – 9
200 – 500	9 – 10
500 – 1000	10 -11
1000 – 5000	11 – 13
5000 – 50000	13 – 17
50000 <	17 – 20

Histogram optymalny



Hazard

Zdecydowana większość gier losowych opiera się na prawdopodobieństwie zdarzenia...

...najprostszy,
jak rzut monetą, ...





...złożony, jak rozdanie pokera...



...całkowicie losowy jak ruletka...

...oraz może być pod tym kątem analizowana.

- **Prawdopodobieństwo trafienia „oczka”**
- **Ilość unikatowych rozdań w pokerze**

Definicja klasyczna prawdopodobieństwa

Każdemu zdarzeniu losowemu A przypisujemy liczbę P(A), zwaną prawdopodobieństwem tego zdarzenia, taką że 0 ≤ P(A) ≤ 1.

» *(Kolmogorov, 1933)*

Prawdopodobieństwem zdarzenia A nazywamy stosunek liczby n_a zdarzeń sprzyjających zdarzeniu A do liczby wszystkich zdarzeń N

$$P(A) = \frac{n_a}{N}$$

$$P(\Omega) = 1$$

Przy czym A jest podzbiorem tzw. zdarzenia pewnego Ω.

$$A \subset \Omega$$

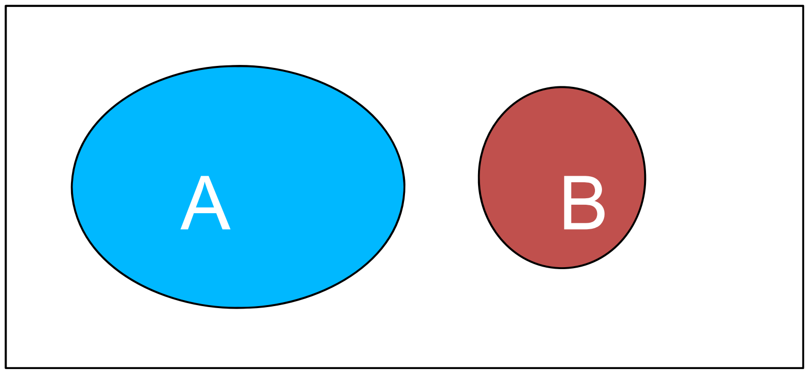
Definicja klasyczna prawdopodobieństwa

Prawdopodobieństwo sumy wzajemnie wykluczających się zdarzeń losowych A i B jest równe sumie prawdopodobieństw tych zdarzeń

» *(Kolmogorov, 1933)*

czyli:

$$P(A \cup B) = P(A) + P(B), \, gdzie \quad A \cap B = \varnothing$$



Przykład zdarzenia losowego

Rzucamy monetą dwa razy. Możliwe wyniki to:

- (o, o) – wyrzucenie dwóch orłów
- (o, r) – wyrzucenie orła, a potem reszki
- (r, o) – wyrzucenie reszki, a potem orła
- (r, r) – wyrzucenie dwóch reszek

Zatem zbiór: E={(o, o); (o, r) ; (r, o); (r, r)}

jest zbiorem zdarzeń elementarnych.

Jeżeli zbiór zdarzeń elementarnych ma E ma n-elementów to zdarzeń losowych jest 2^n

Przykład zdarzenia losowego

W tej sytuacji możliwych jest 2^4 zdarzeń losowych.

Niektóre zdarzenia losowe, np.:

A = {(o,o); (o,r); (r,o)} – wyrzucenie co najmniej 1 orła

B = {(o,o); (o,r);} - orzeł w pierwszym rzucie

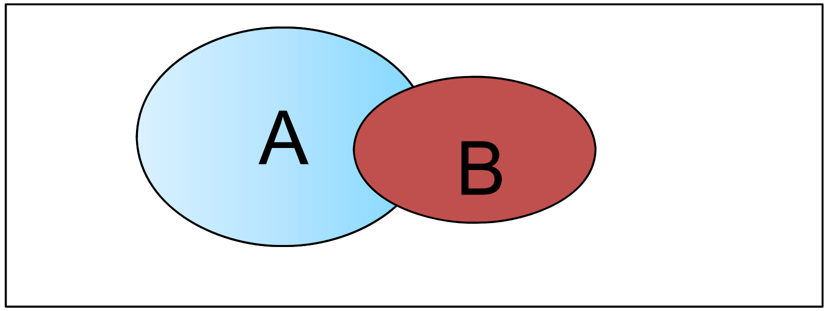
G = {(o,o);} - wyrzucenie dwóch orłów

H = {(o,r); (r,o)} – wyrzucenie dokładnie jednej reszki

Relacje zdarzeń

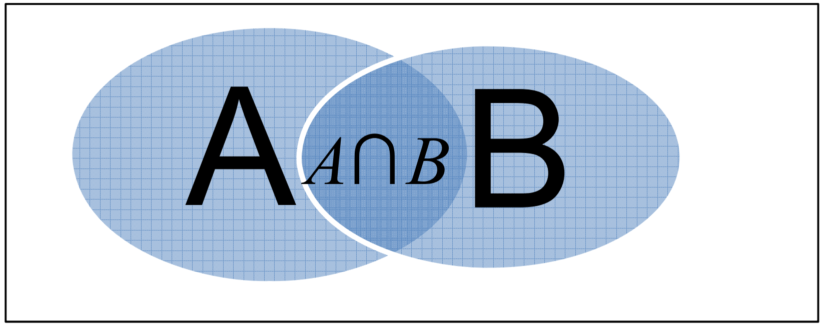
Suma zdarzeń – zachodzi co najmniej jedno ze zdarzeń A,B

$$A \cup B$$



Iloczyn zdarzeń – zachodzi zdarzenie A oraz zdarzenie B

$$A \cap B$$



Relacje zdarzeń

Zdarzenie przeciwne – <u>nie zachodzi</u> zdarzenie A

$$A'$$

Zdarzenie A pociąga zdarzenie B (operator: zbiór A zawiera się w zbiorze B)

$$A \subset B$$

Zdarzenia A i B wykluczające się

$$A \cap B = \varnothing$$

Rzut kostką

Rzut kostką: wynik od 1 do 6 oczek.

Prawdopodobieństwo wyrzucenia czterech oczek:

1/6 (teoria)



Rzut kością

<u>3 rzuty:</u> 4, 5, 4	(wyrzucenie 4: 2/3 >> 1/6)
<u>10 rzutów:</u> 4, 4, 5, 3, 2, 6, 1, 2 , 2, 4	(wyrzucenie 4: 3/10 > 1/6)
<u>1000 rzutów:</u> 4,5,4,3,1,5,1,2,1,3,2,2,6,6,5,4,...	(wyrzucenie 4: ~1/6)

Częstość zdarzenia zmierza do wartości prawdopodobieństwa dopiero przy wielokrotnym powtórzeniu.

Ogólna (aksjomatyczna) definicja prawdopodobieństwa.

Zakładamy, że zdarzenia losowe A i B są podzbiorami tego samego zbioru zdarzeń elementarnych E.

Def. **Jeśli każdemu zdarzeniu losowemu A przyporządkowano liczbę rzeczywistą P(A) zwaną prawdopodobieństwem zdarzenia A, w taki sposób, aby spełnione były następujące warunki:**

I

$$0 \le P(A) \le 1$$

II prawdopodobieństwo zdarzenia pewnego jest równe 1

$$P(E)=1$$

III prawdopodobieństwo sumy dwóch wykluczających się zdarzeń jest równe sumie prawdopodobieństw tych zdarzeń

$$P(A \cup B) = P(A) + P(B), gdzie \quad A \cap B = \varnothing$$

to określoną w ten sposób funkcję P nazywamy prawdopodobieństwem

Prawdopodobieństwo warunkowe

Ogólna definicja:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

Przy czym P(B) > 0 (tj. zdarzenie B musi być jakkolwiek prawdopodobne)

Efektywnie, każde prawdopodobieństwo jest warunkowe: np. dla danych zdarzeń A i B:

$$P(A) = P(A \mid \Omega)$$

$$A \cap B = \varnothing \Rightarrow P(A \mid B) = 0$$

$$A \subset B \Rightarrow P(A \mid B) = \frac{P(A)}{P(B)}$$

$$B \subset A \Rightarrow P(A \mid B) = 1$$

Kostka – przykład 1

Rzucamy trzema kostkami 6-cio-ściennymi. Wiemy, że na każdej kostce wypadła inna liczba oczek. Jakie jest prawdopodobieństwo że na jednej kostce wypadło 5 pod warunkiem że na każdej kostce wypadła inna liczba?

$$P(A \cap B) = \frac{5 \cdot 4 \cdot 3}{\Omega}$$

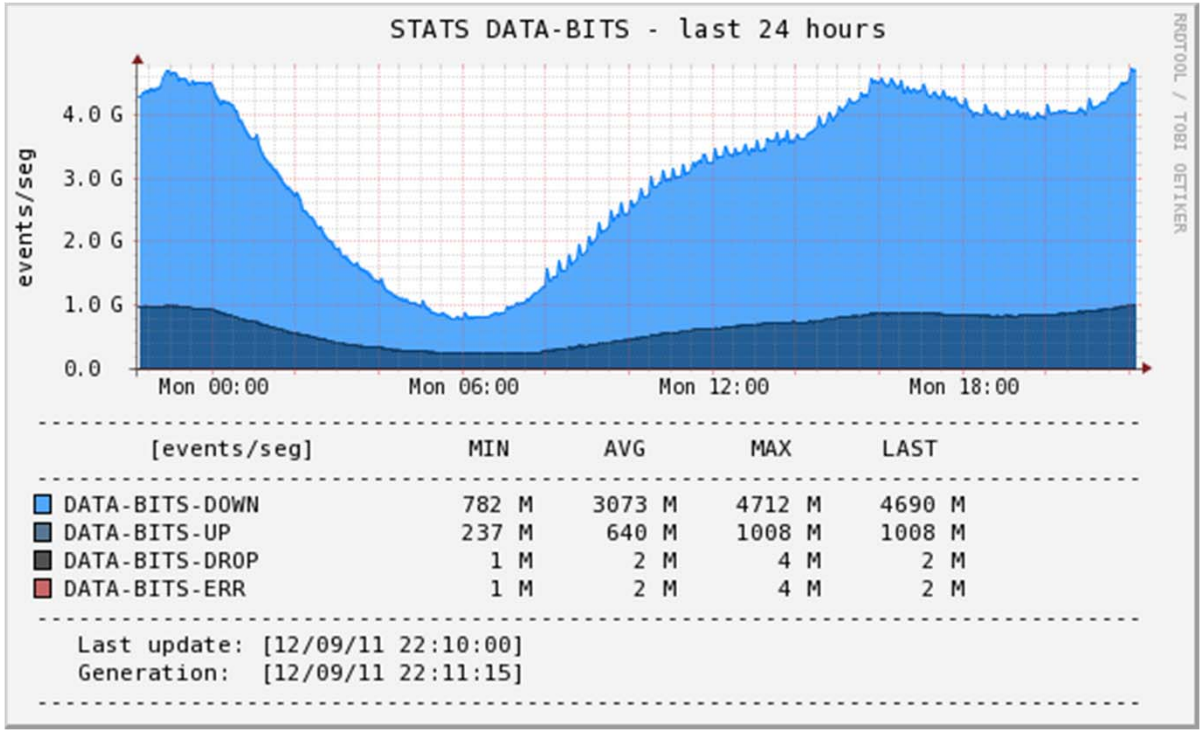
$$P(B) = \frac{6 \cdot 5 \cdot 4}{\Omega}$$

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{5 \cdot 4 \cdot 3 \Omega}{6 \cdot 5 \cdot 4 \cdot \Omega}$$

Telekomunikacja

Obciążenie sieci telekomunikacyjnej w ciągu dnia:

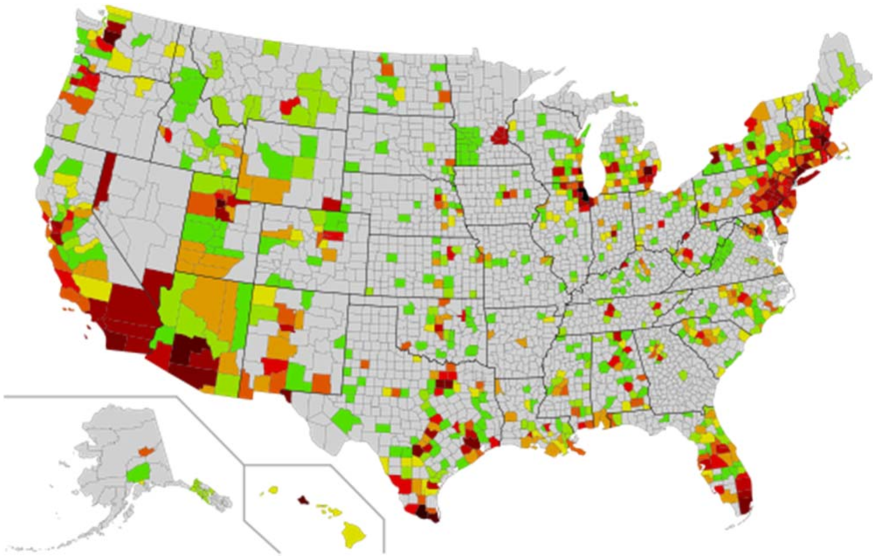
- **przesyłanie danych w Internecie**
- **ilość wykonanych połączeń telefonicznych**



(Źródło: http://rrdtool.cs.pu.edu.tw/gallery/index.en.html)

Epidemiologia

Statystyka umożliwia analizę i modelowanie rozwoju chorób oraz pomaga zapobiegać epidemiom.



Liczba zachorowań na świńską grypę w roku 2009 w USA
(Źródło: http://commons.wikimedia.org)

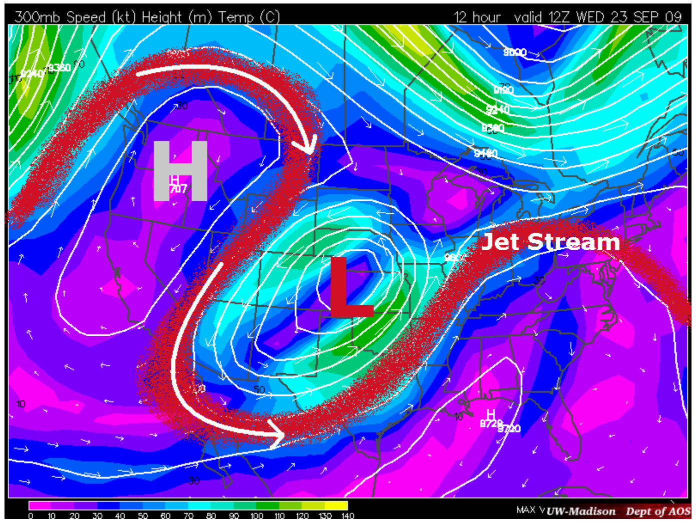
Statystyka medyczna, np. średnia liczba zachorowań w regionie

Statystyka społeczna, np. gęstość zaludnienia

Statystyka gospodarcza, np. PKB, wydatki na opiekę zdrowotną

Meteorologia

Modele pogodowe umożliwiające przewidywanie pogody oraz <u>wykrywanie potencjalnych kataklizmów</u>, np. huraganów



(Źródło:stormdebris.net/Math_Forecasting.html)

